AHUMADA: A LARGE SPEECH CORPUS IN SPANISH FOR SPEAKER IDENTIFICATION AND VERIFICATION

J. Ortega-García[1], J. González-Rodríguez[1], V. Marrero-Aguiar[2], Cg. J. J. Díaz-Gómez[3], Cap. R. García-Jiménez[3], Cap. J. Lucena-Molina[3], Tcol. J. A. G. Sánchez-Molero[3]

[1] DIAC, EUIT- Telecomunicación, Univ. Politécnica de Madrid
28031 Madrid, Spain. e-mail: jortega@diac.upm.es
[2] Dpt. Lengua Española, Univ. Nacional de Educación a Distancia
[3] Servicio de Policía Judicial

ABSTRACT

Speaker Recognition is a major task when security applications through speech input are needed. Regarding speaker identity, several factors of variability must be considered: a) Factors concerning peculiar intra-speaker variability (manner of speaking, inter-session variability, dialectal variations, emotional condition, etc.) or forced intra-speaker variability (Lombard effect, cocktail-party effect). b) Factors depending on external influences (kind of microphone, channel effects, noise, reverberation, etc). To cope with all these variability sources, a specific speech database called AHUMADA has been designed and collected for speaker recognition tasks in Castilian Spanish. AHUMADA incorporates six different recording sessions, including both *in situ* and telephone speech recordings. A total of 104 male speakers uttered isolated digits, digit strings, phonologically balanced short utterances, phonologically and syllabically balanced read text and more than one minute of spontaneous speech, so about 15 GB of speech material is available. Speaker verification results, concerning the available variability sources are also presented.

1. INTRODUCTION(*)

As it has been already mentioned, speech variability is a main degradation factor in speaker recognition tasks. Both intra-speaker and external variability sources produce mismatch between training and testing phases. Usually, training phase is accomplished under controlled or supervised conditions, referred as (ideal) "laboratory" conditions, while testing phase is done under unpredictable or even unknown real conditions. This mismatch between phases causes appreciable degradation in recognition experiments, and many robust techniques have been proposed to deal with it [1, 2]. Anyway, in many cases, the mismatch problem is still an open question.

Our goal in this paper is to present AHUMADA speech database, the first large corpora in Castilian Spanish designed for speaker recognition purposes, in the line of other existing large corpora for other languages [3, 5, 6]; and then, testing how some of the variability factors included in it may affect speaker verification experiments. Some examples of the variability factors included in AHUMADA corpus can be: *in situ* recordings and telephone speech; read texts at different speech rate; read speech versus spontaneous speech; different microphones and telephone handsets; inter-session variability in six different recording sessions; dialectal variations of speakers (which may be even different for one particular speaker when reading or naturally speaking); or fixed utterances for all speakers through all sessions versus specific utterances for each speaker in each session.

The paper is organized as follows. In section 2, the design and collection of AHUMADA speech corpus is presented. In section 3, the speaker verification system used over the speech corpus available is described. In section 4, some results at testing phase, concerning the available variability sources present in it, are shown. Finally, some conclusions are extracted and some future work is proposed in section 5.

2. 'AHUMADA' SPEECH DATABASE

2.1. Design of the Speech Corpus

The speech corpus has been designed to include many of the speaker variability sources, allowing us to focus on them and study their underlying effects in speaker verification systems [4]. In this sense, the enrolled speakers where requested to utter the following:

- a) 24 isolated digits, discarding the first and the last two of them due to prosodic considerations. The remaining 20 digits consist in two repetitions of isolated digits from 0 to 9.
- b) 10 digit strings consisting of ten digits each, being the first five strings identical for all speakers through all recording sessions, and the last five strings specific for each speaker for all sessions.
- c) 10 phonologically balanced utterances of 8-12 word length. These utterances were identical for all speakers through all sessions.
- d) 1 phonologically and syllabically balanced text, of about 180 words (more than 1 minute of duration), read

(*) *This work has been supported by CICYT under Project TIC97-1001-C02-01*

at a normal speaking rate. This text was fixed for all speakers through all sessions.

- e) 2 repetitions of the previous fixed text, asking the speakers to read it at a fast and at a slow speaking rate. (this task was only requested in sessions 1, 3 and 5, where *in situ* recordings were accomplished).
- f) 1 specific text, different from speaker to speaker and from session to session, for each speaker. This text was randomly selected from novels and newspapers, and at least 1 minute of this kind of speech is available.
- g) More than 1 minute of spontaneous speech, asking every speaker to describe (without long pauses and hesitations) whatever they wanted. There were available some paintings and pictures, and subjects like "describe your last holidays", "describe the place where you live/were born", etc., were also suggested.

2.2 Phonological and Syllabic Balance

Tasks 2.1.c) and 2.1.d) have been specifically designed in order to reproduce the frequency of appearance of phonemes and syllabic schemes in spoken Castilian Spanish [8]. The selected lexicon corresponds to the most usual in Spanish [9]. The 'standard' frequency of appearance used in the design phase has been measured over an oral corpus of more than 20,000 words [10].

The total number of phonemes in task 2.1.c) is 409. The correlation coefficient (Pearson test) between Spanish 'standard' phonological appearance and the designed utterances was 0.9966. In the same task, the total number of syllables was 185 with a syllabic correlation coefficient of 0.9963. In task 2.1.d), a fixed text for all speakers with about 180 words, there is a total number of phonemes of 712. The correlation coefficient between Spanish 'standard' phonological appearance and the designed text was 0.9988. Moreover, the total number of syllables in it was 305, being in this case the correlation coefficient 0.9960. In both tasks, the level of significance is 0.001 (the maximum attainable).

2.3 Data Collection and Recording Sessions

As it has been previously mentioned, six recording sessions were established. Sessions 1, 3 and 5 were *in situ* recorded in a quiet room and supervised by a trained operator. In each of these *in situ* recordings, two different input channels were simultaneously used: in one of them, the same microphone was used for all sessions; in the other, different microphones were used from session to session.

The notation used to specify both microphones in each case is MIC*n*_1 and MIC*n*_2, were *n* corresponds to one of the three possible sessions. Consequently, MIC1_1, MIC3_1 and MIC5_1 were the same microphone, namely SONY ECM-66B, lavalier unidirectional electret type, at about 10 cm. from the speaker mouth. MIC1_2 is an AKG D80S dynamic cardioid microphone, placed on a desk at about 30 cm. from speaker. MIC3_2 is an AKG C410-B head-mounted dynamic microphone. MIC5_2 is a low-cost Creative Labs desk microphone for PC sound-card applications.

In sessions 2, 4 and 6, telephone line was used to collect the data. In session 2, every speaker was making a phone call from the same telephone in an internal-routing call. In session 4, speaker were requested to make a local call from its own home telephone, trying to search a quiet environment (they were asked to be alone in a closed room). In session 6, a local call was made from a quiet room, using 10 different standard handsets [7]. In this last telephone recording session, simultaneous microphone acquisition was performed (MIC6_2), using the same lavalier type SONY microphone as in MIC1_1, MIC3_1 and MIC5_1.

In each session, both microphones (connected through a high-quality Behringer MIC502 preamplifier) and telephone lines (connected through a specific adapter) were fed to a professional DAT device (Tascam DA-30 MKII), were digital recording at 44.1 kHz. was accomplished.

2.4 Recording-Room Acoustics

A quiet room was selected to make the recordings of sessions 1, 3, 5 and 6 (simultaneous telephone and microphone input). No anechoic chamber or acoustic cabin was used, as we wanted to have real-environment (quiet) recording conditions. To avoid undesired room reverberation, several acoustic panels were placed around the desk were recordings were made.

Measurements done with acoustic specific equipment showed good acoustic conditions for the speech recording sessions. An equivalent noise level of only 27 dBA was measured, and the upper limit for the reverberation time in a third-octave band analysis is 0.48 secs.

2.5 Distribution of Ages

The distribution of ages was designed in order to model a possible distribution of users of a certain speaker recognition application. Equi-distribution of ages may not respond to real users distribution, and more weight has been applied to the range of ages between 28 and 42 years. Figure 1 shows the distribution in five-year periods.

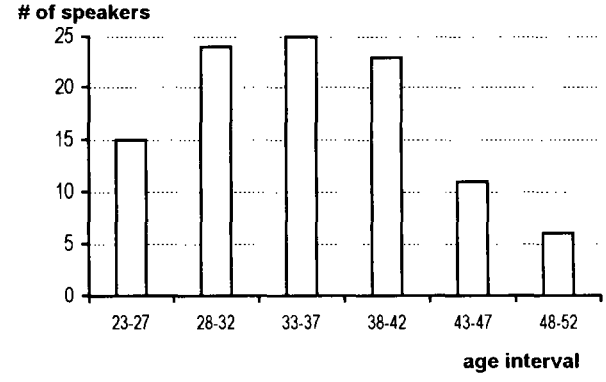

Fig. 1.- Distribution of ages in five-year intervals for AHUMADA speech corpus.

2.6 Time Interval between Sessions

As inter-session variability is an important factor to be taken into account in speaker recognition-oriented databases, at least a time interval separation of 15 days between equivalent sessions (on one side, microphone sessions 1, 3 and 5, and on the other side, telephone sessions 2, 4 and 6) was meant to. Anyway, the enrollment availability of speakers may have caused some deviations from these initial requirements.

Recordings began in June 1997, with session 1. Following, it can be found the time intervals between session 1 and the rest of the sessions:

- *Session 2:* 73% of recordings were done within 15 days interval from session 1. Specifically, 36% were accomplished the same day of session 1. The maximum time interval (100% of recordings) is 40 days.
- *Session 3:* 80% of recordings were done between 20 and 40 days after session 1, and 92% between 15 and 50 days. The minimum interval is 10 days after session 1, and the maximum is 80 days.
- *Session 4:* 73% of recordings were accomplished in a time interval of 15 to 50 from session 1. 19% were done between 40 and 80 days after session 1.
- *Session 5:* The minimum interval between session 1 and session 5 is 30 days. 77% of them were acquired between 40 and 80 days after session 1, while 10% were separated in time from 80 to more than 90 days.
- *Session 6:* The minimum time interval of session 6 recordings is 30 days after session 1. 78% of speech material was recorded between 40 and 80 days after session 1. The last 9% of recordings were done between 80 and more than 90 days after session 1.

3. SPEAKER VERIFICATION SYSTEM

In order to perform some speaker recognition tests over the available data, a speaker verification system has been used. As we wanted to evaluate text-independent verification results, Gaussian Mixture Models (GMM) [11, 15] have been used. Due to the lack of time, tests were accomplished over a subset of (randomly-selected) 25 speakers from the total number of 104 available speakers. As a previous stage, silences longer than 0.8 s. were removed, and the first 40 s. of speech have been used for training purposes. Read fixed text (task 2.1.d) from session 1 has been used to train in all cases the system, generating one model per speaker. All speech material used for training and testing has been down-sampled to 8 kHz. Cepstral coefficients derived from LPC analysis (LPCC) of order 10 have been used as feature vectors. Frames of 30 ms. taken every 120 samples with Hamming windowing and pre-emphasis factor of 0.97 are used as input to the system. As in some cases there was not enough speech material for the testing phase, overlapping between consecutive testing sequences has been forced: 0% for 5 s. sequences, 50% for 10 s. sequences and 66.6% for 15 s. sequences.

All 25 speakers were used as claimants for their corresponding models and as imposters for the rest of speaker models. Tests without normalization and with likelihood-domain normalization [12, 13, 14] have been accomplished. As the density at point X (input sequence) for all speakers other than the true speaker, S, is frequently dominated by the density for the nearest reference speaker, we have applied the following normalization criterion:

$$\log L(X) = \log p(X|S = S_c) - \max_{S \in \text{ref}, S \neq S_c} \log p(X|S)$$

where S_c means claimed speaker model. Balance between false rejection error and false alarm errors is searched, so equal error rate (EER) for each speaker is computed, and average EER through all speakers for each case is presented in the next section.

4. SPEAKER VERIFICATION RESULTS

As it has been already mentioned in the preceding section, model training has been performed using about 40 s. of read speech from a fixed text, equal for all speakers (task 2.1.d), corresponding to session 1 and using MIC1_1. The remaining speech from this task (same session, same microphone) has been used for initially testing the verification system, in order to establish some baseline results for the rest of testing experiments. Figure 2 shows these results.

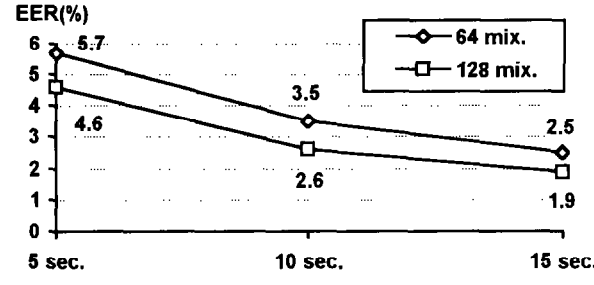

Figure 2.- Verification results with both 64 mixture and 128 mixture GMMs when no normalization is applied.

Baseline results in Figure 2 do not include normalization. When likelihood-domain normalization was applied, EERs less than 0.5% were found in all referred cases. Figure 3 shows verification results when testing was accomplished with spontaneous speech (task 2.1.g) from session 1 using MIC1_1.

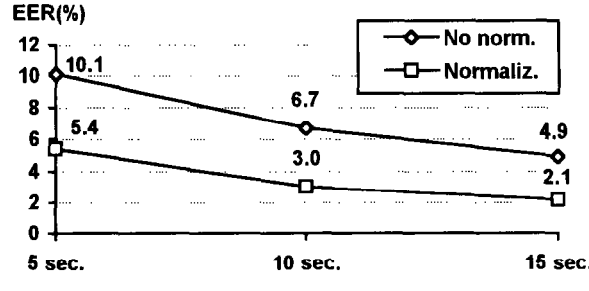

Figure 3.- EER for different duration of testing sequences of spontaneous speech, session 1, MIC1_1.

In Figure 4, same training text used for training and testing (task 2.1.d) of session 1, but considering the effect of using the second microphone (MIC1_2).

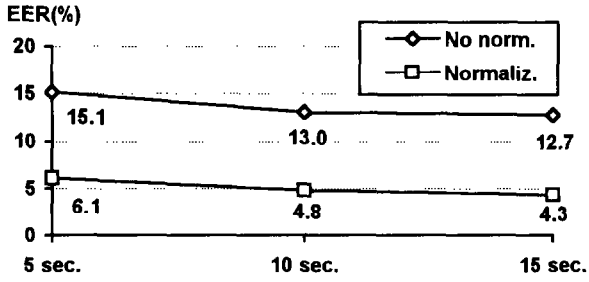

Figure 4.- Verification results due to using different microphones in training and testing phases.

Finally, Figure 5 presents EER for testing sequences of spontaneous speech (task 2.1.g) of session 5 with MIC5_1.

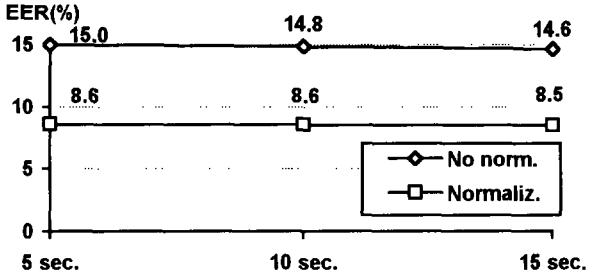

Figure 5.- Inter-session variability between sessions 1 and 5, testing with spontaneous speech.

5. CONCLUSIONS AND FUTURE WORK

A large Castilian Spanish corpus for speaker recognition tasks has been presented. Speaker verification experiments described in Section 4 show excellent results when same session, same microphone, same task, and enough amount of testing data (15 s.) is used: normalizing the results of Fig. 1 gives less than 0.5% EER. These two last mentioned parameters, namely testing sequence lenght and likelihood-domain normalization, produce, with no doubt, significant improvements in all cases. When just the kind of speech is changed, from read speech to spontaneous descriptive speech (Fig. 3), EER increases to (in the best case) 2.1% which is still an acceptable limit. Nevertheless, if we use read speech to test but we change the microphone used (Fig. 4) we get a best EER of 4.3%. If we focus on inter-session variability (Fig. 5) with spontaneous testing speech, 8.5% EER is obtained as best.

Anyway, these results may only give a certain initial idea of the possibilities that AHUMADA database can offer in speaker recognition tasks. In this sense, the use of more efficient features, including Δ and $\Delta\Delta$ cepstra, Λ and $\Lambda\Delta$ energy; the use of channel compensating techniques like CMN and RASTA; the use of multi-session and multi-task training; the use of more sophisticated normalization schemes, including general population models, etc., and the testing results for all 104 speakers, will focus the work to be done in the near future over the multi-variability data of AHUMADA corpus.

ACKNOWLEDGEMENTS

The authors wish to thank David Pérez-Alonso and Carlos Bravo-Ruíz who did the hard work of obtaining the speaker verification results on AHUMADA database (in such a short period of time!).

6. REFERENCES

[1] Acero A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Dordrecht (NL), 1993.

[2] Junqua J.-C. and Haton J.-P., *Robustness in Automatic Speech Recognition -Fundamentals and Applications*, Kluwer Academic Publishers, Dordrecht (NL), 1996.

[3] Godfrey J., Graff D. and Martin A., "Public Databases for Speaker Recognition and Verification", *ESCA Workshop on Automatic Speaker Recognition*, pp. 39-42, Martigny (CH), April 94.

[4] Gibbon D., Moore R. and Winski R., eds., *Handbook of Standards and Resources for Spoken Language Systems*, EAGLES Spoken Language Working Group, Mouton de Gruyter, 1997.

[5] Naik J., "Speaker Verification over the Telephone Network: Databases, Algorithms and Performance Assessment", *ESCA Workshop on Automatic Speaker Recognition*, pp. 31-38, Martigny (CH), April 94.

[6] Boves L. *et al.*, "Design and Recording of Large Data Bases for Use in Speaker Verification and Identification", *ESCA Workshop on Automatic Speaker Recognition*, pp. 43-46, Martigny (CH), April 94.

[7] Reynolds D., "HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects", *IEEE Intl. Conf. on Acous. Speech and Signal Proc. ICASSP-97*, pp. 1535-1542, Munich (D), April 97.

[8] Guerra R., "Recuento Estadístico de la Sílaba en Español", *Estudios de Fonética*, 1, pp. 9-112: Collectanea Phonetica, VII, CSIC, Madrid (E), 1983.

[9] Juilland A. and Chang-Rodríguez E., *Frequency Dictionary of Spanish Words*, Mouton, The Hague (NL), 1969.

[10] Quilis A. and Esgueva M., "Frecuencia de Fonemas en el Español Hablado", *LEA*, 2, pp. 1-25, 1980.

[11] Reynolds D., *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph. D. Thesis, Georgia Institute of Technology, 1992.

[12] Furui S., "An Overview of Speaker Recognition Technology", *ESCA Workshop on Automatic Speaker Recognition*, pp. 1-9, Martigny (CH), April 94.

[13] Matsui T. and Furui S., "Similarity Normalization Method for Speaker Verification Based on A Posteriori Probability", *ESCA Workshop on Automatic Speaker Recognition*, pp. 59-62, Martigny (CH), April 94.

[14] Rosenberg A. E., DeLong J., Lee C.H., Juang B. H., and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification", *Proceedings of Intl. Conf. on Spoken Language Proc., ICSLP-92*, pp. 599-602, Banff (Canada), 1992.

[15] J. Ortega-García and J. González-Rodríguez, "Providing Single- and Multi-Channel Acoustical Robustness to Speaker Identification Systems ", *IEEE Intl. Conf. on Acous. Speech and Signal Proc., ICASSP-97*, pp. 1107-1110, Munich (D), 1997.